

Как организовать поиск в стартапе, который планирует вырасти до масштабов ВКонтакте

Богдан Гаркушин

Руководитель поискового направления ВКонтакте







20 лет в поиске

Когда-нибудь найду)

Разрабатывал поиск, был менеджером, руководил продуктом и отвечаю за поисковое направление

Богдан Гаркушин

Руководитель поискового направления ВКонтакте

О чем мы сегодня поговорим

1

Движки

- База данных
- Lucene/Vispa/Spinx/..

2

Ранжирование

- Формула вручную
- МЛ
- Сложный МЛ

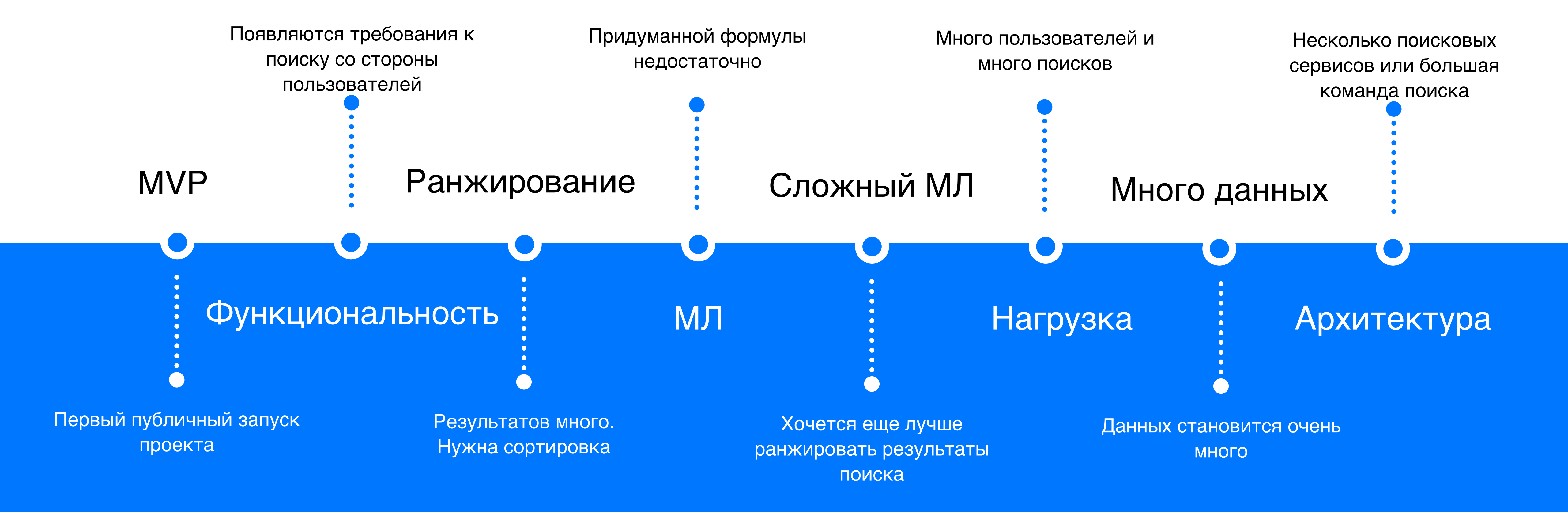
3

Архитектура

- Репликация
- Шардирование

Стадии развития вашего проекта

Движки | Ранжирование и ML | Архитектура



Метрики качества поиска

Аудиторные

- DAU/WAU/MAU
- Поисковые сессии

Поисковые

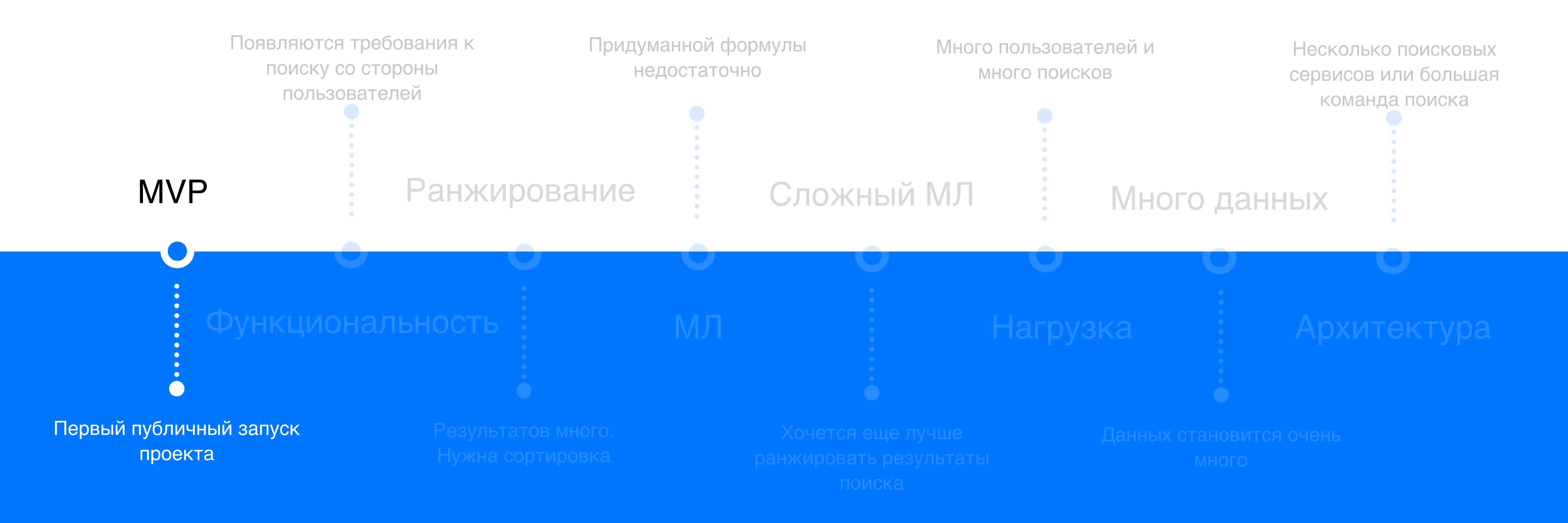
- Поиски с результатами
- Поиски с кликами

Конверсионные

- Поиски с действиями

Подготовка к MVP

Движки | Ранжирование и ML | Архитектура



Подготовка к MVP

Продукт не запустили

Данных мало

Результатов поиска мало

Аудитории нет

Нагрузки нет

Все оценки на бумаге

Как вы представляете себе поиск?

дима



Дима Трундуков

Санкт-Петербург

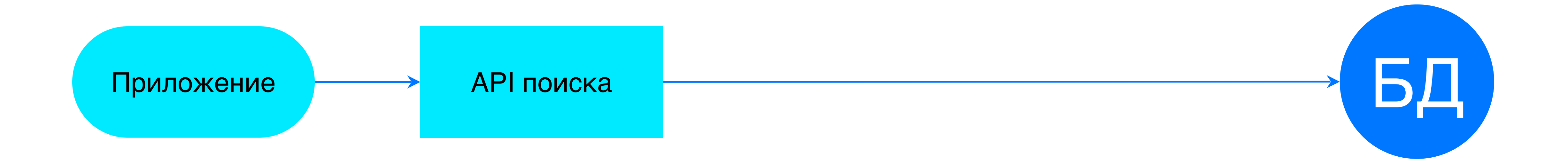
ВКонтакте

Написать сообщение · Позвонить

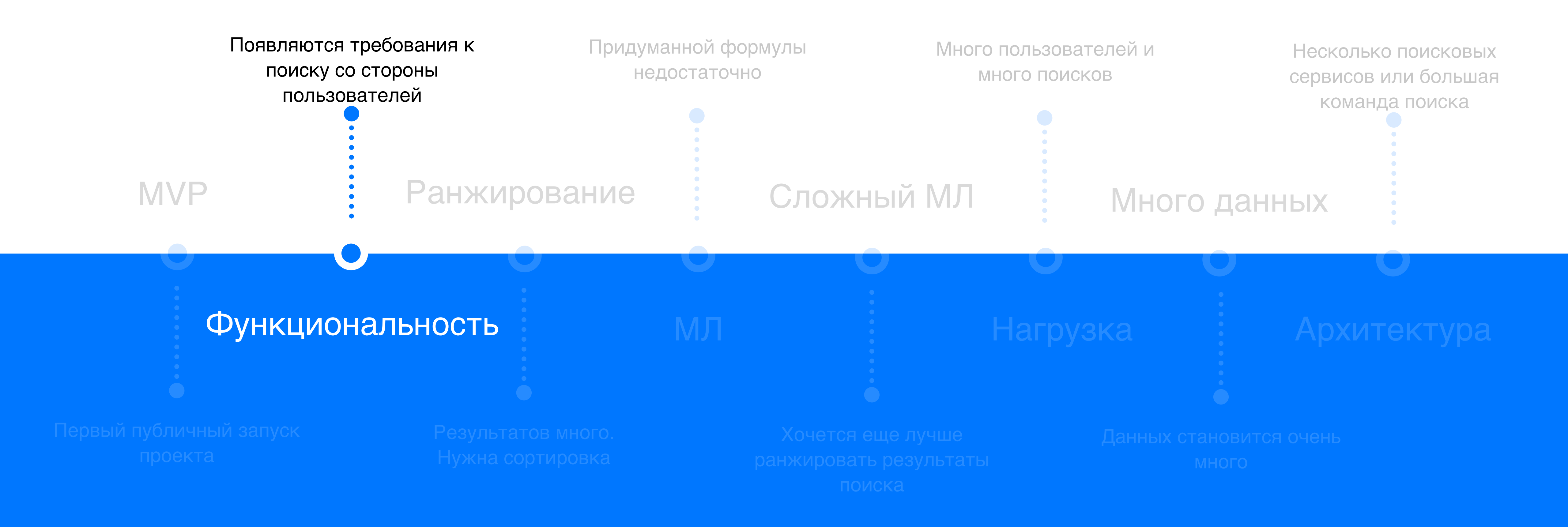
Удалить из друзей

Разрабатываем MVP и готовимся запуску

Все максимально просто. Работа нацелена на максимальную простоту



Проект запустили и появились пользователи




Продукт запустили
Появились первые функциональные требования
Данных мало
Первые пользователи
Нагрузка маленькая
Результатов поиска мало
Аудитория небольшая

Что нужно изменить в результатах?

дима



Дима Трундуков
Санкт-Петербург
ВКонтакте
Написать сообщение · Позвонить

Удалить из друзей



Dima Orlov
Санкт-Петербург
СПбГУТ им. Бонч-Бруевича
Написать сообщение

Добавить в друзья

Что нужно изменить в результатах?

дима  



Дмитрий Костенко
Москва
НИУ ВШЭ (ГУ-ВШЭ)
Написать сообщение · Позвонить

Удалить из друзей



Dmitry Viazmin
Написать сообщение

Добавить в друзья

На что повлияют наши изменения?

Аудиторные

- DAU/WAU/MAU
- Поисковые сессии

Поисковые

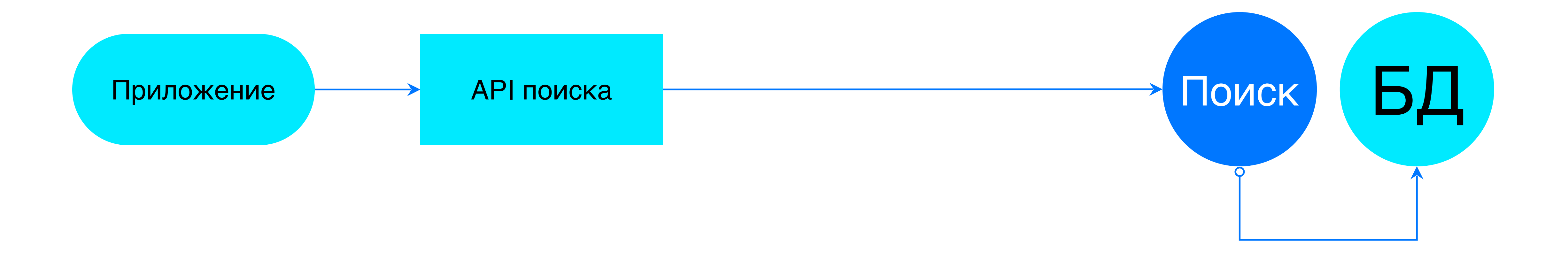
- **Поиски с результатами**
- Поиски с кликами

Конверсионные

- Поиски с действиями

Появляется поисковый движок

Поиск обеспечивается за счет движка, который регулярно синхронизирует свое состояние с базой данных сервиса



Sphinx

+ интеграция с СУБД

+ API для php, Python, Java



VESPA

+ векторный поиск

+ ядро на C++

+ горизонтальное масштабирование

+ поддержка МЛ

+/- ориентирован на Kubernetes



ElasticSearch, SOLR

+ Готовый движок на базе Lucene
+ Быстрый старт

- Отстают от Lucene
- Проблемы при больших объемах

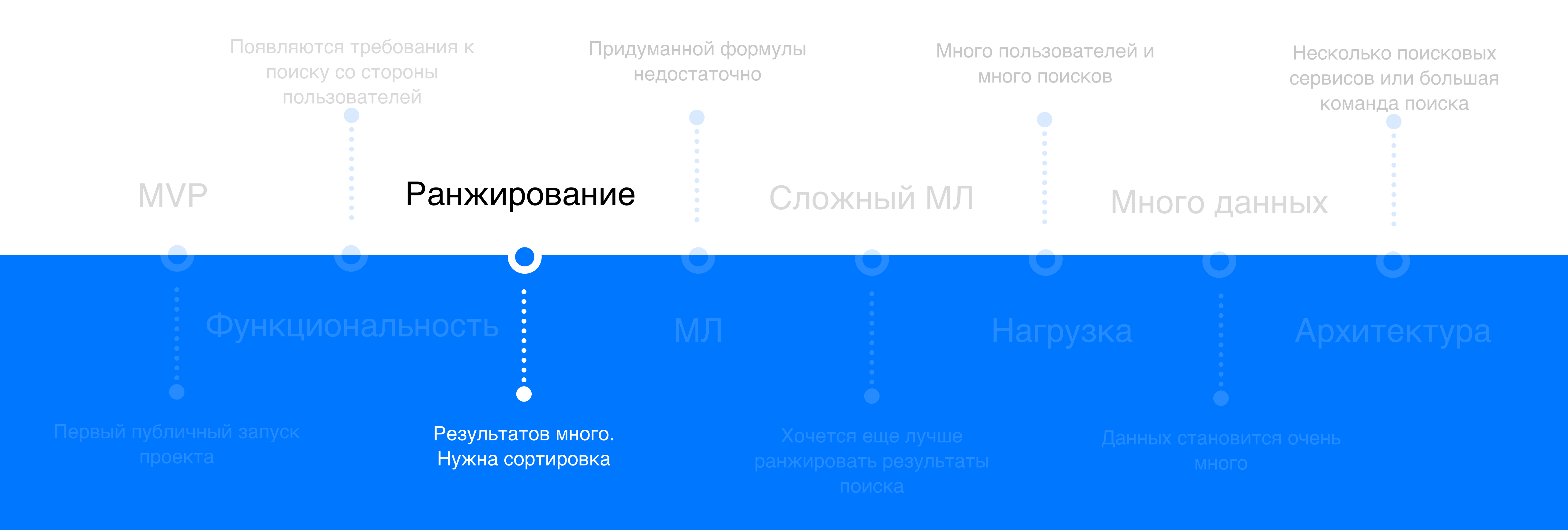
Lucene

+ Популярная библиотека поиска
+ Полнотекст, нечеткий поиск и …
+ Расширяется и кастомизируется
+ Оптимизирован под современные процессоры

- Это движок, поиск делать самостоятельно

Результаты не помещаются на экран устройства

Движки | **Ранжирование и ML** | Архитектура



Продукт
запустили

Функциональные
требования

Данных много

Много
результатов
поиска

Нагрузка
маленькая

Результаты поиска
не помещаются на
экран устройства

Аудитория
небольшая

Результатов больше 10

Люди 4 107

алена кузьмина





Алёна Кузьмина
Санкт-Петербург
Школа кундалини йоги Амрит Нам Саровар СПб
Написать сообщение · Позвонить

Удалить из друзей



Алена Кузьмина
Москва
Российское представительство в ООН
Написать сообщение

Добавить в друзья



Алёна Кузьмина
Санкт-Петербург
Дом Астролога
Написать сообщение

Добавить в друзья

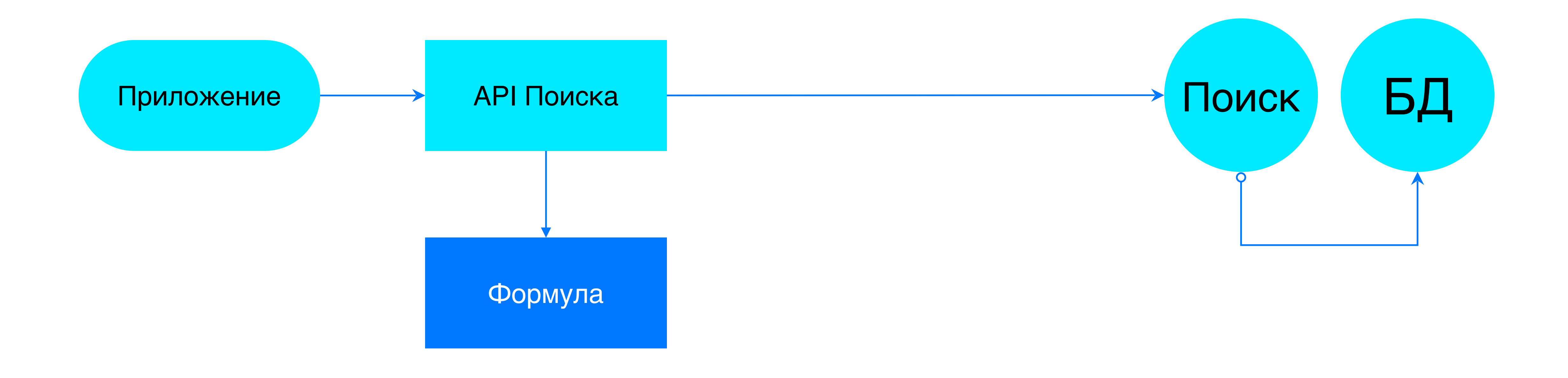


Алёна Кузьмина
Москва
МГУ
Написать сообщение

Добавить в друзья

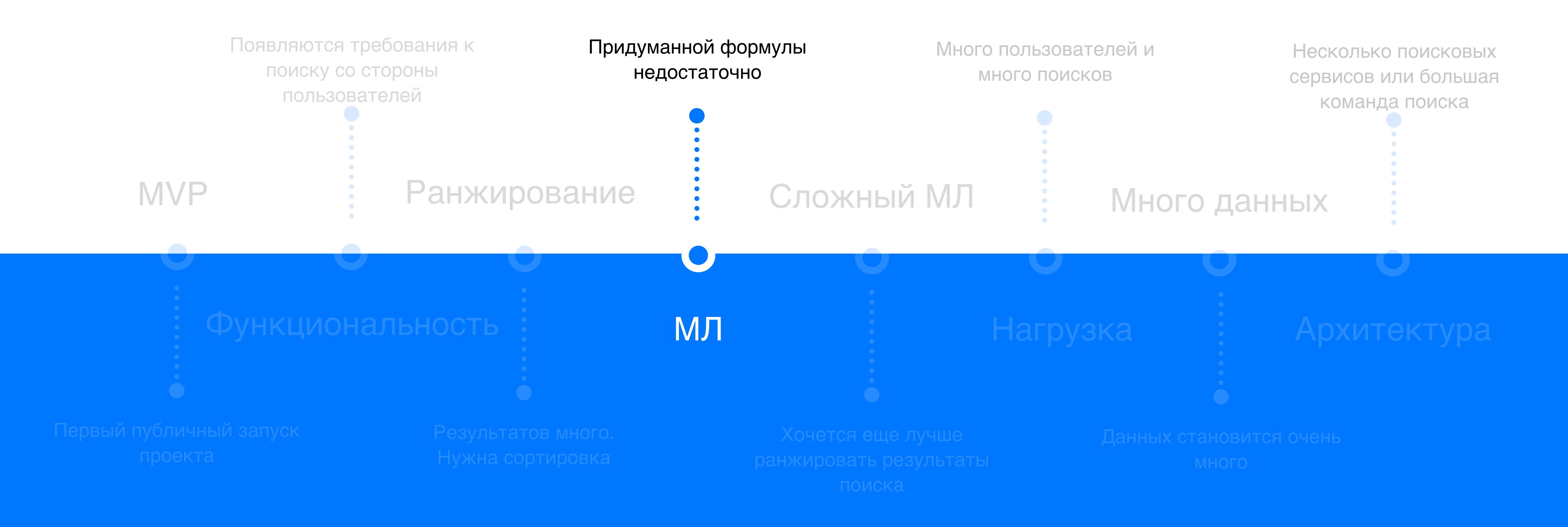
Формула

Мы все достаточно квалифицированны или считаем себя такими, чтобы точно сказать, что нужно пользователю, и придумать для него формулу ранжирования.



Результатов поиска очень много

Движки | **Ранжирование и ML** | Архитектура



Продукт запустили

Функциональные требования

Данных много

Очень много результатов поиска

Нагрузка маленькая

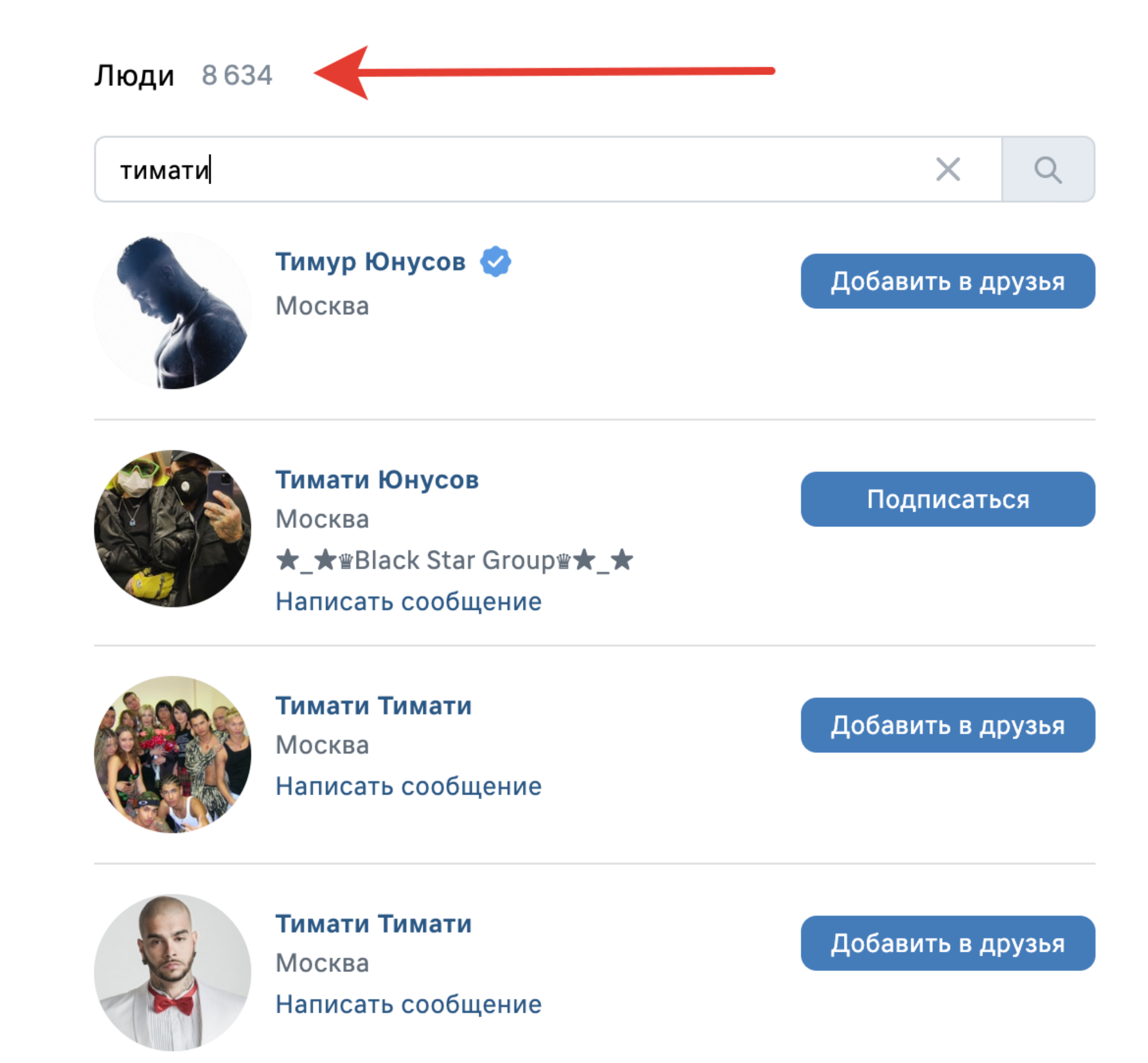
Результатов поиска очень много

Аудитория небольшая

Результатов очень много

Критерии

- Звезды
- Друзья
- Новые знакомые
- Соц. граф



На что повлияют наши изменения?

Аудиторные

- DAU/WAU/MAU
- Поисковые сессии

Поисковые

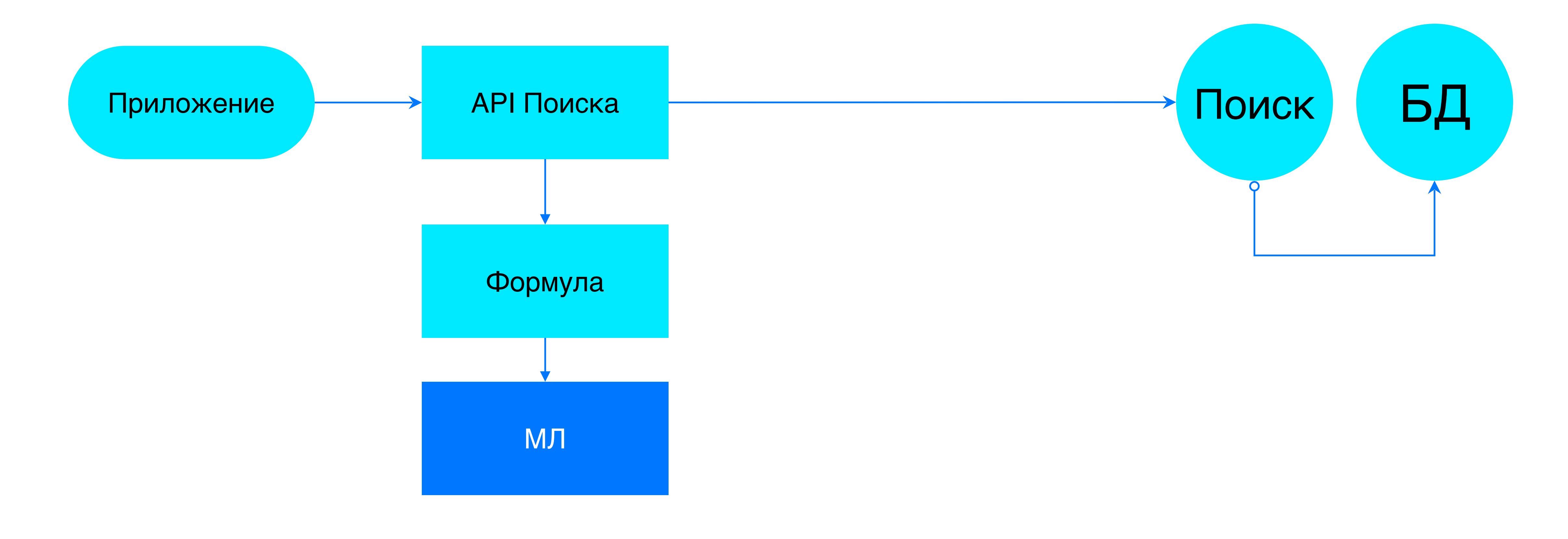
- Поиски с результатами
- Поиски с кликами

Конверсионные

- Поиски с действиями

Машинное обучение

Машинное обучение - это онлайн и оффлайн процесс. В оффлайне мы работаем с историческими данными и подбираем формулу, в онлайне мы собираем то, что знаем здесь и сейчас и ранжируем согласно подобранной ранее формуле



Хотим лучше понимать пользователя

Движки | **Ранжирование и ML** | Архитектура

Появляются требования к поиску со стороны пользователей

Придуманной формулы недостаточно

Много пользователей и много поисков

Несколько поисковых сервисов или большая команда поиска

MVP

Ранжирование

Сложный МЛ

Много данных

Функциональность

МЛ

Нагрузка

Архитектура

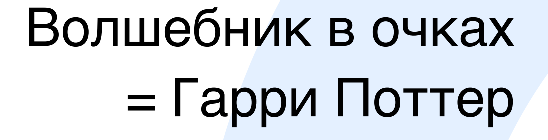
Первый публичный запуск проекта

Результатов много. Нужна сортировка

Хочется еще лучше ранжировать результаты поиска

Данных становится очень много

Лучше понимать пользователя

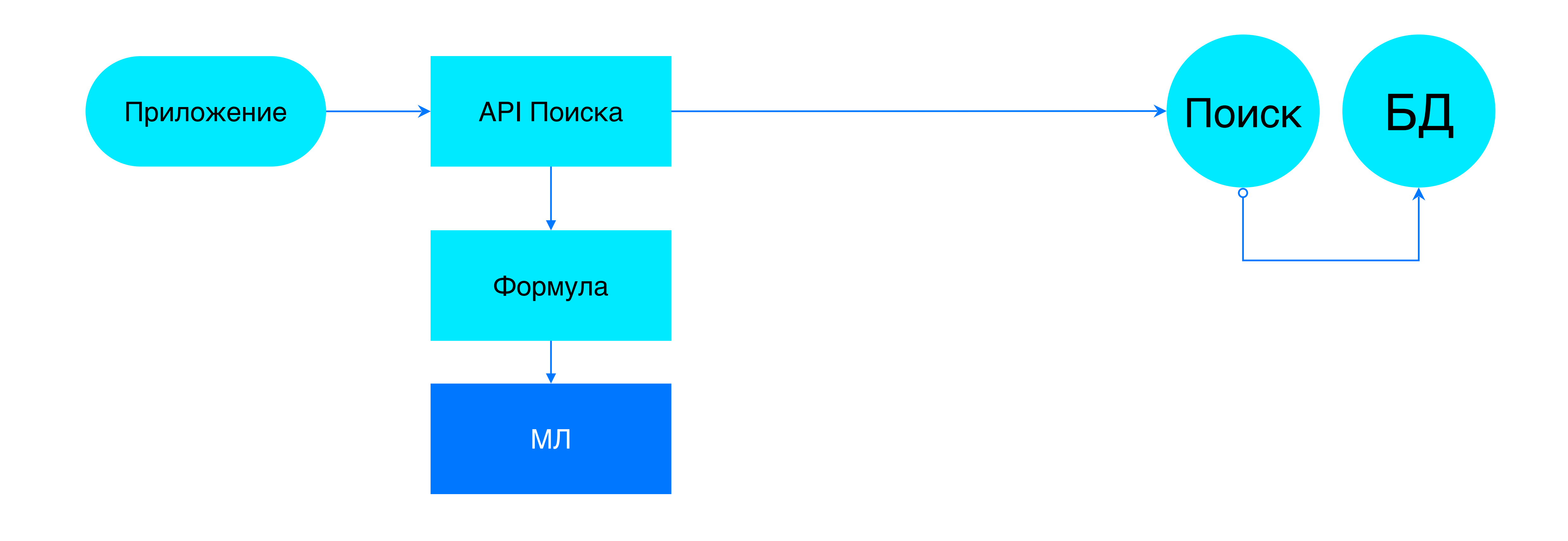


Комик в очках =
Гарик Харламов

Ведущий КВН =
Александр Масляков
Дмитрий Хрусталев
…

Где искать новые
источники для
синонимов?

Машинное обучение



Рост аудитории -> рост нагрузки

Движки | Ранжирование и ML | **Архитектура**

Появляются требования к поиску со стороны пользователей

Придуманной формулы недостаточно

Много пользователей и много поисков

Несколько поисковых сервисов или большая команда поиска

MVP

Ранжирование

Сложный МЛ

Много данных

Функциональность

МЛ

Нагрузка

Архитектура

Первый публичный запуск проекта

Результатов много. Нужна сортировка

Хочется еще лучше ранжировать результаты поиска

Данных становится очень много

Продукт
запустили

Функциональные
требования

Данных много

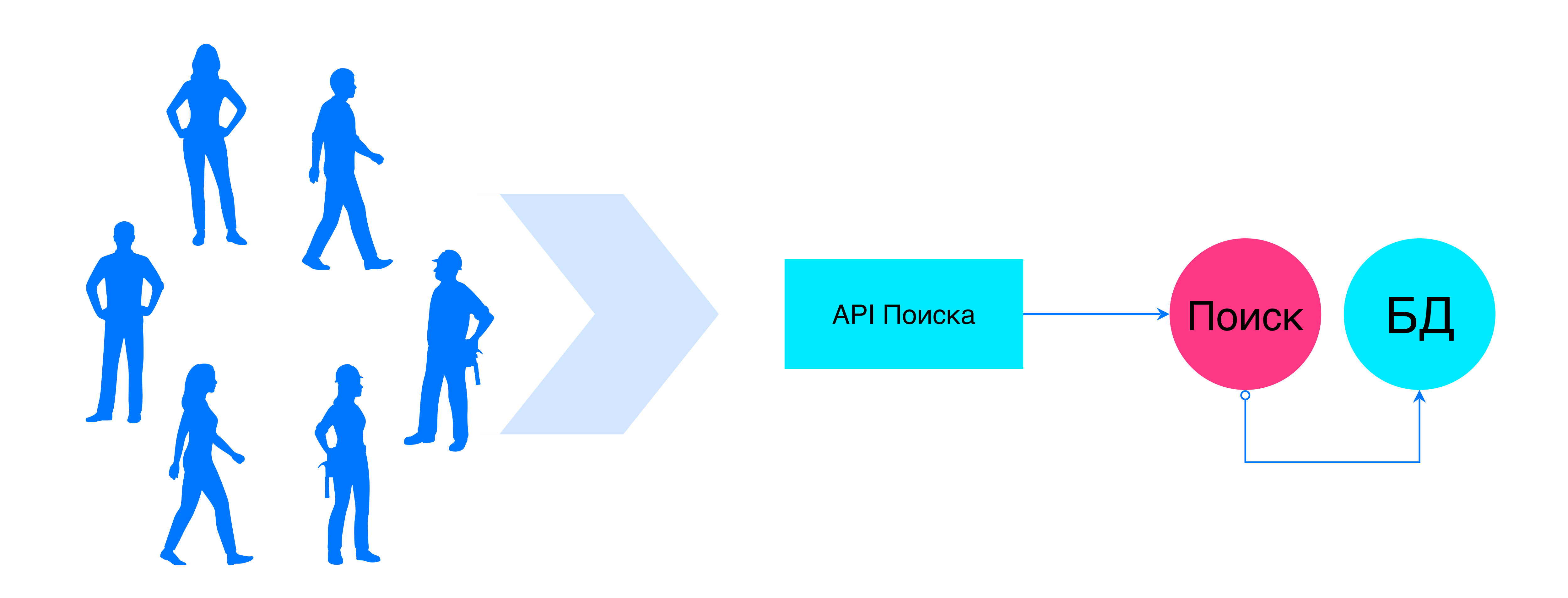
Высокая
нагрузка

Нагрузка
возросла

Результатов
поиска очень много

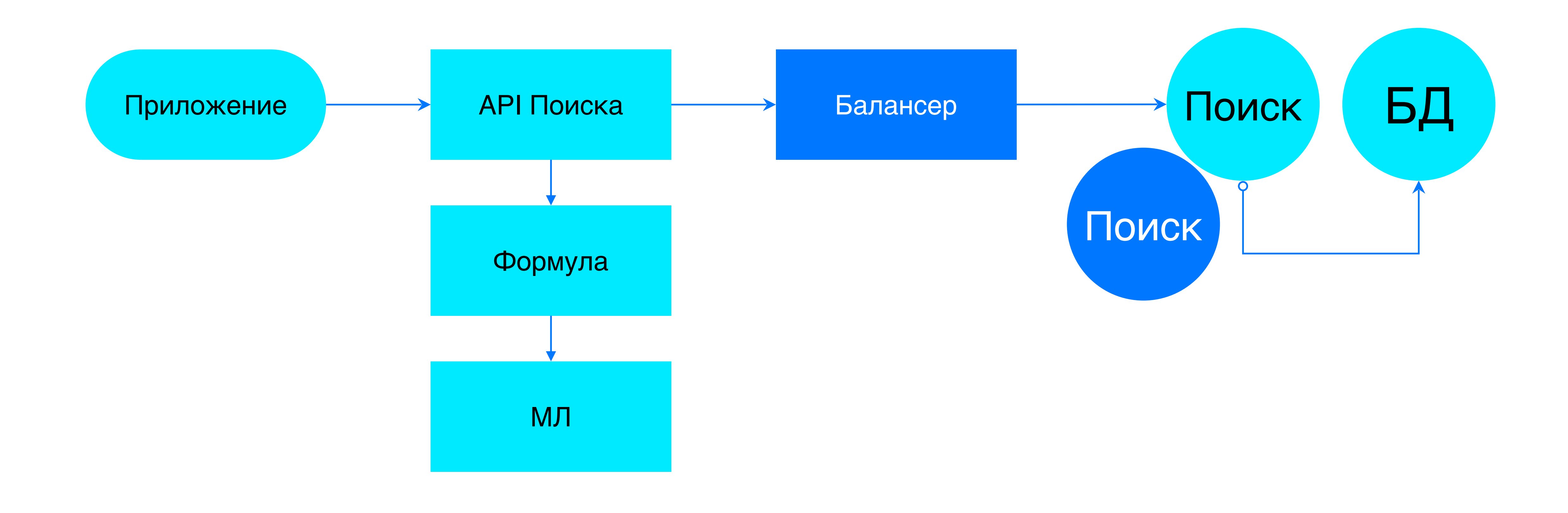
Аудитория
увеличилась

Пользователей становится много



Репликация

Добавляем еще один или несколько абсолютно таких же поисковых серверов



Колоссальный рост базы сервиса

Движки | Ранжирование и ML | **Архитектура**

Появляются требования к поиску со стороны пользователей

Придуманной формулы недостаточно

Много пользователей и много поисков

Несколько поисковых сервисов или большая команда поиска

MVP

Ранжирование

Сложный МЛ

Много данных

Функциональность

МЛ

Нагрузка

Архитектура

Первый публичный запуск проекта

Результатов много. Нужна сортировка

Хочется еще лучше ранжировать результаты поиска

Данных становится очень много

Продукт
запустили

Функциональные
требования

Данные не
помещаются на
один сервер

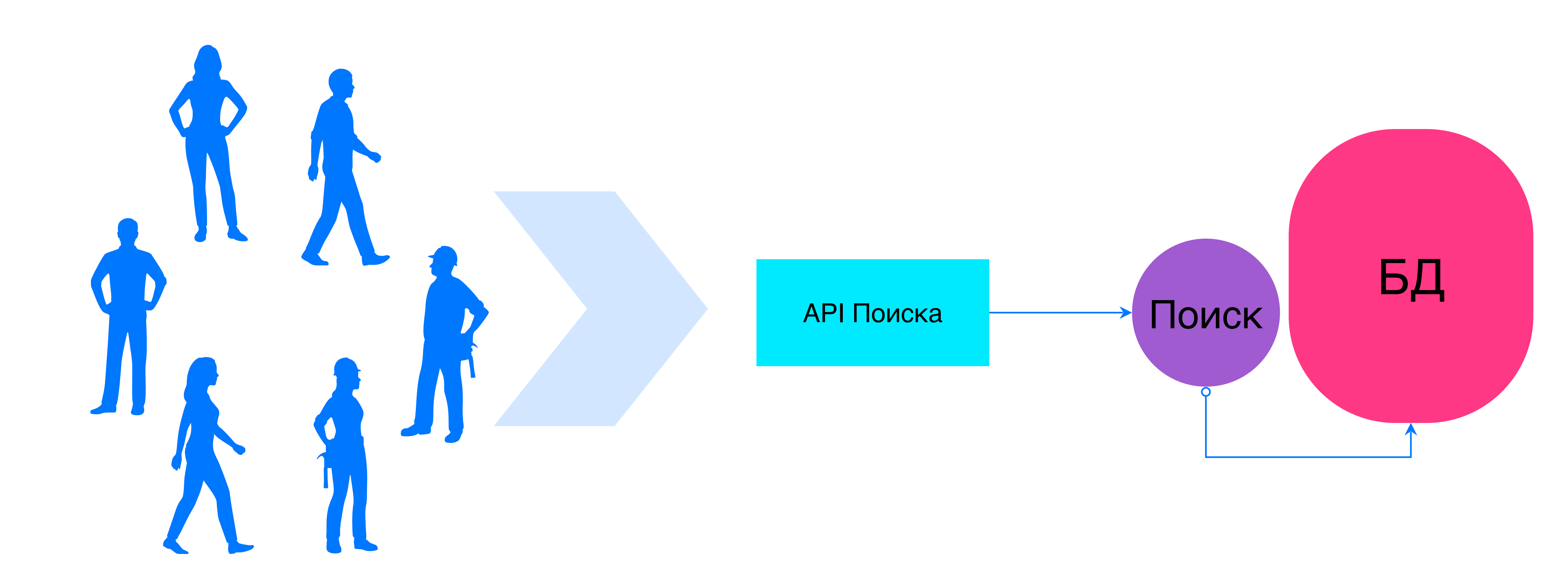
Данных очень много

Нагрузка
возросла

Результатов
поиска очень много

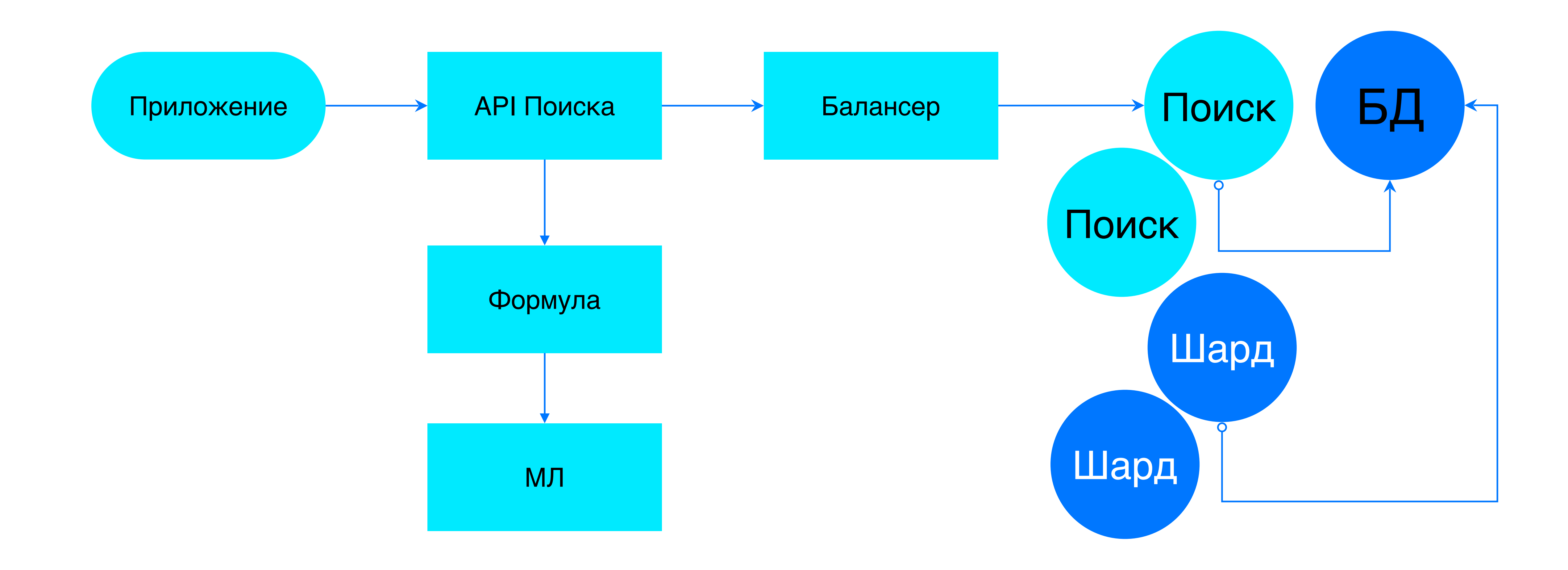
Аудитория
увеличилась

Стало много данных



Шардирование

Добавляем еще один или несколько поисковых серверов с другими данными.



Много разных сервисов

Движки | Ранжирование и ML | **Архитектура**

Появляются требования к поиску со стороны пользователей

Придуманной формулы недостаточно

Много пользователей и много поисков

Несколько поисковых сервисов или большая команда поиска

MVP

Ранжирование

Сложный МЛ

Много данных

Функциональность

МЛ

Нагрузка

Архитектура

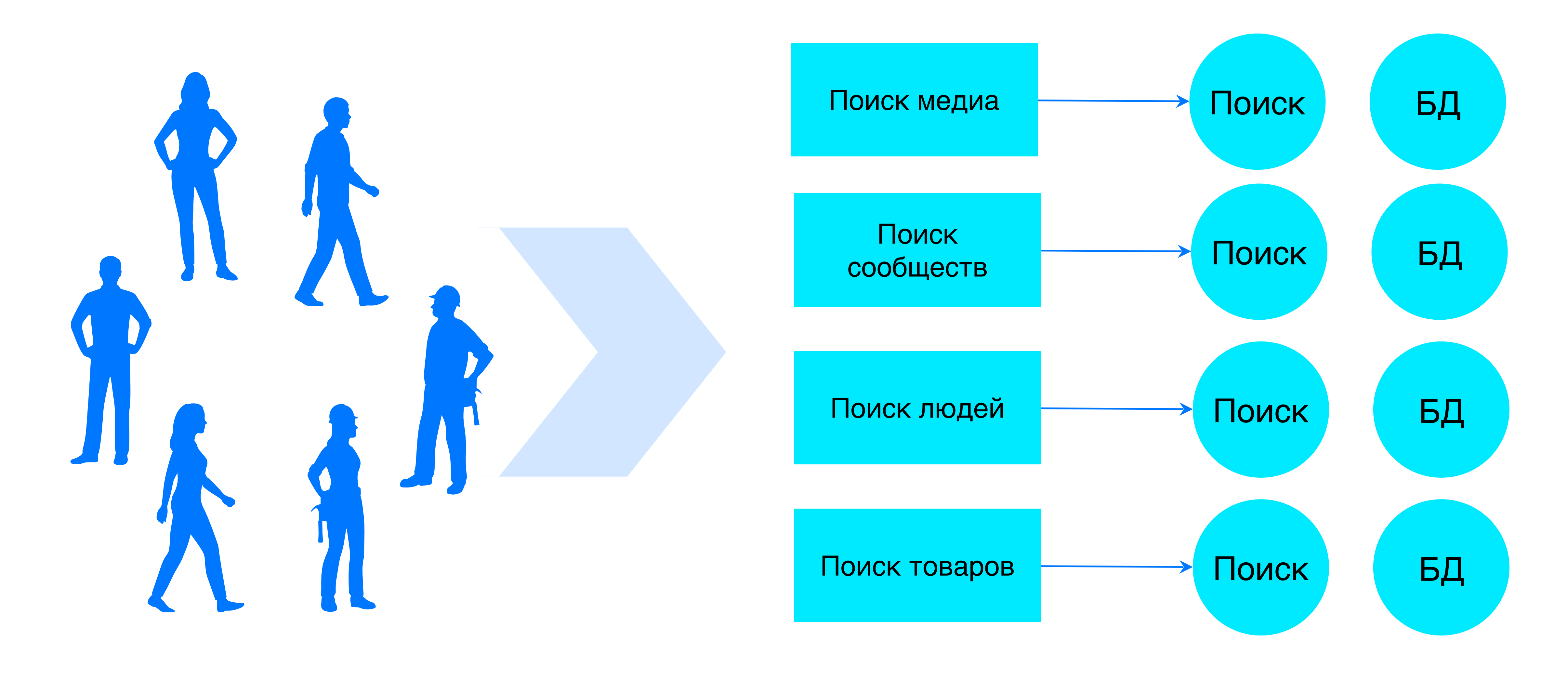
Первый публичный запуск проекта

Результатов много. Нужна сортировка

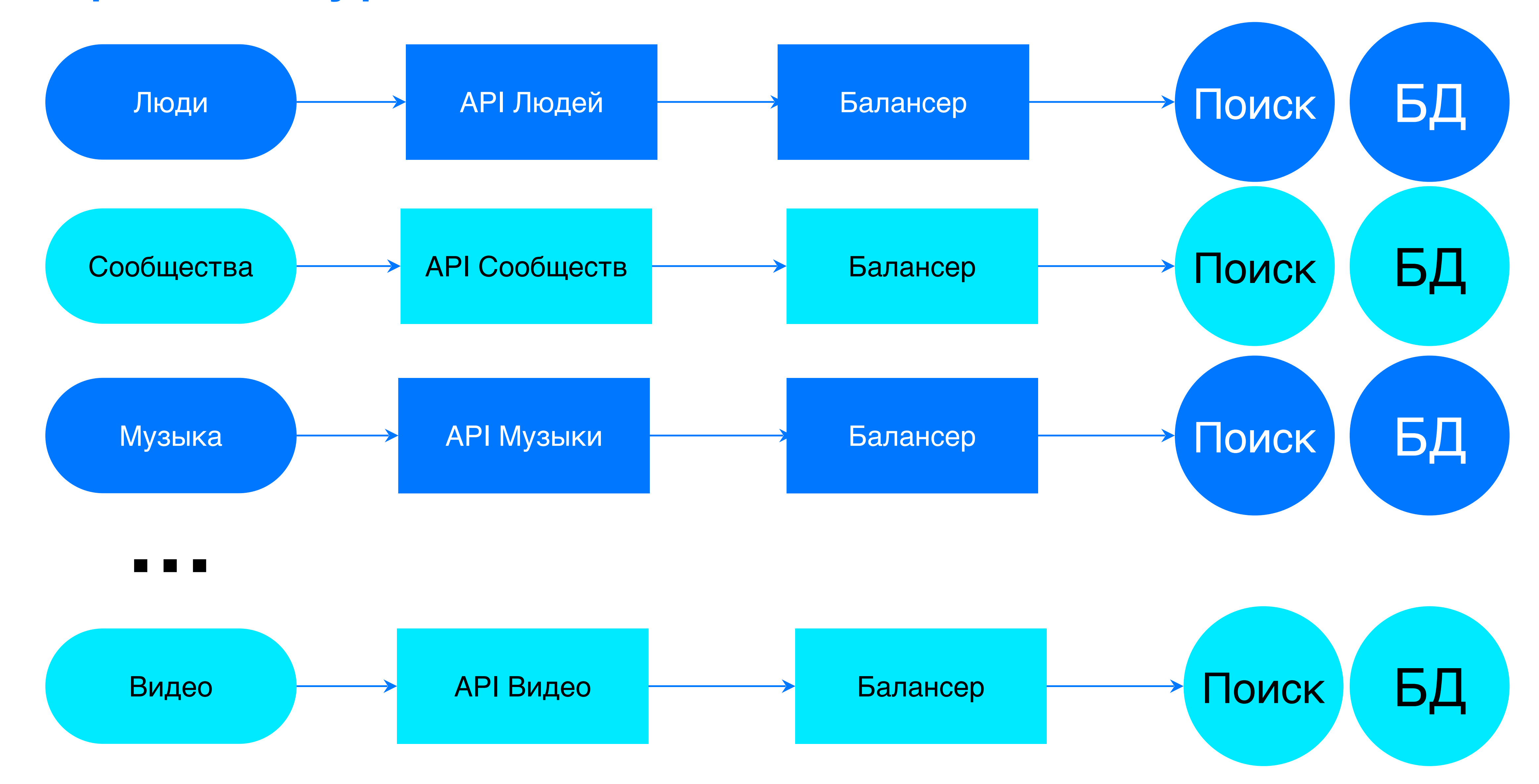
Хочется еще лучше ранжировать результаты поиска

Данных становится очень много

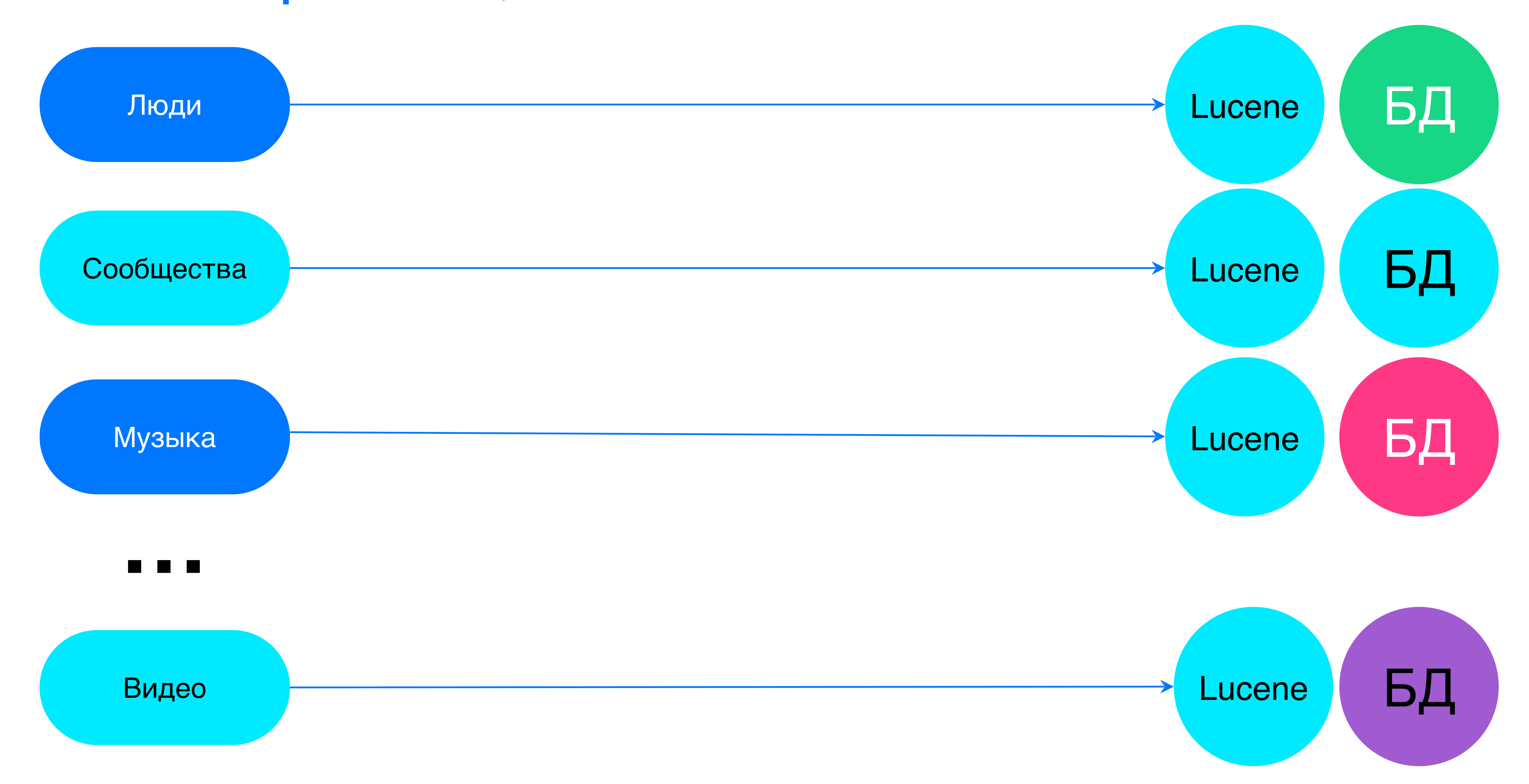
Количество сервисов увеличивается



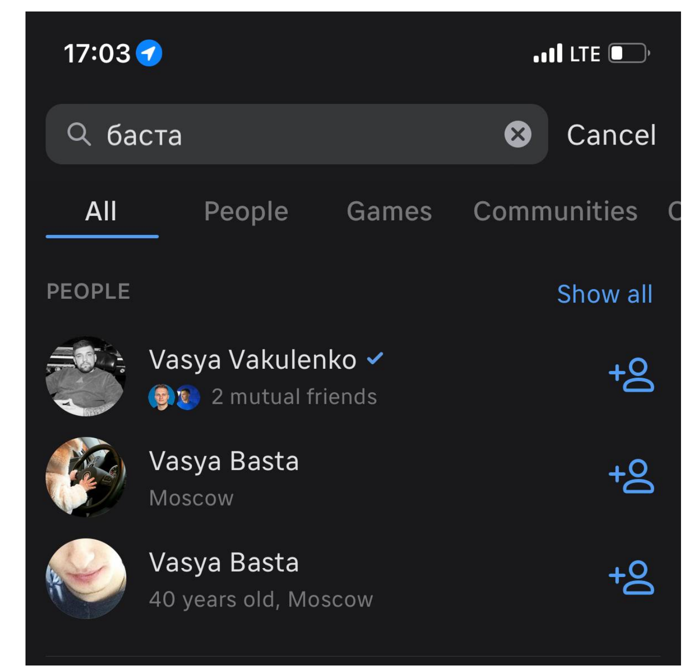
Архитектура поиска ВКонтакте

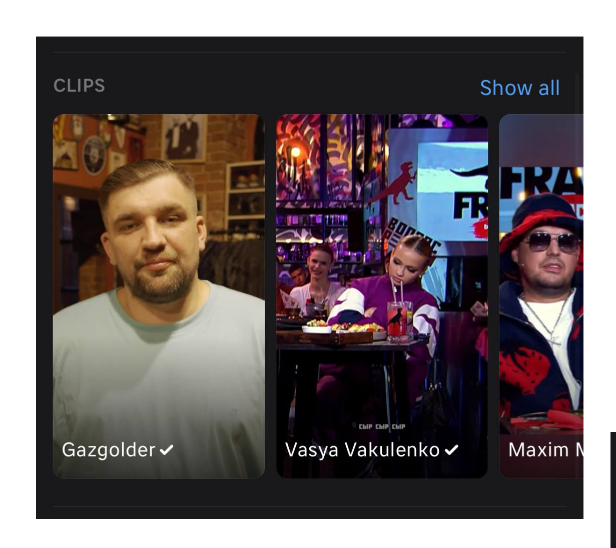


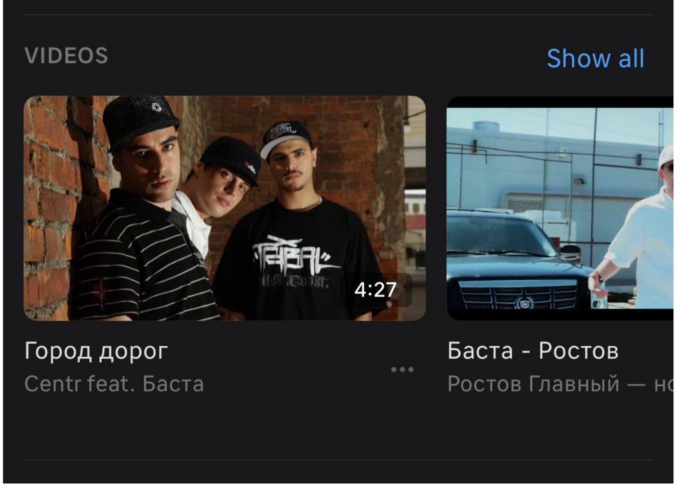
Поиски разные, а движок один

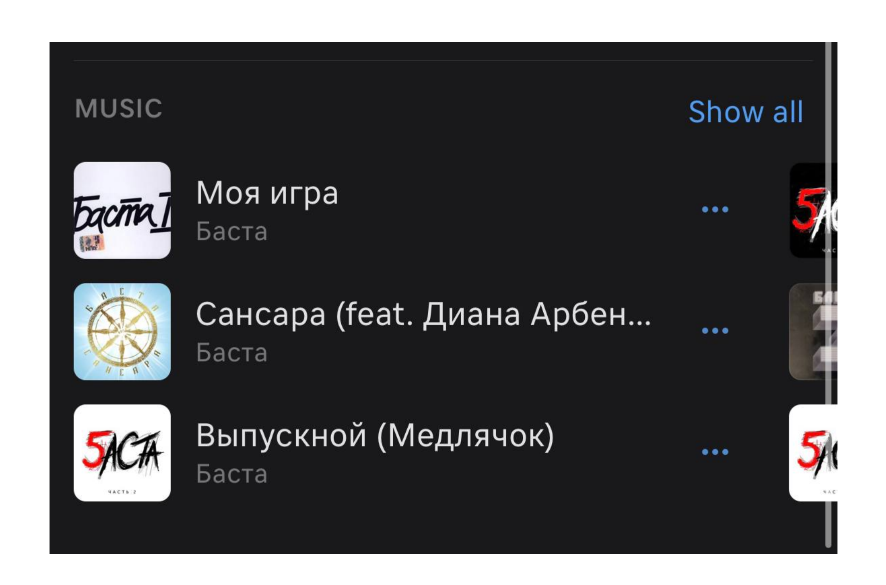


Глобальный поиск











Поиск ВКонтакте сегодня

20 млн
пользователей в день

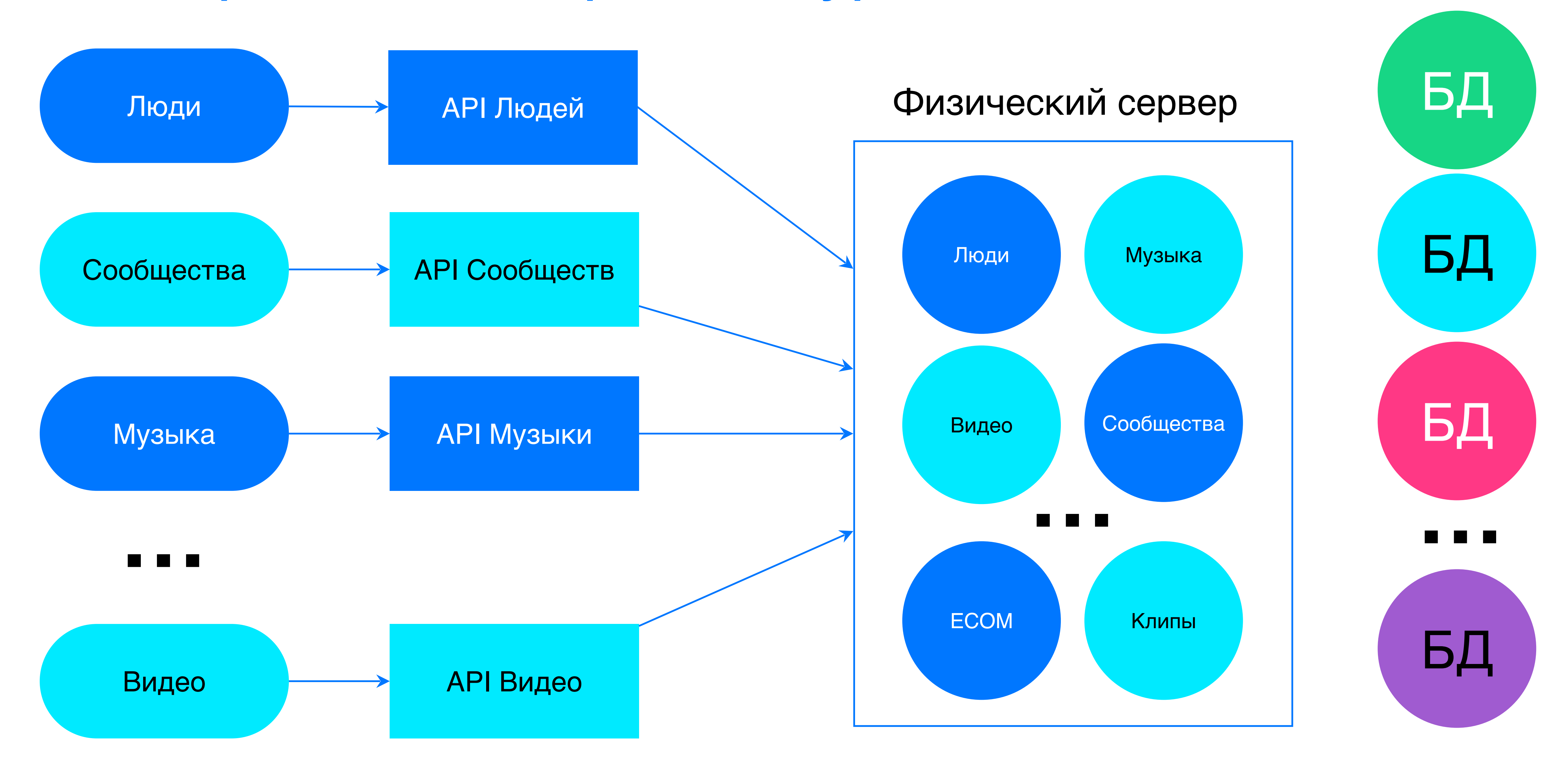
54
поисковых кластера

6
крупных сервисов

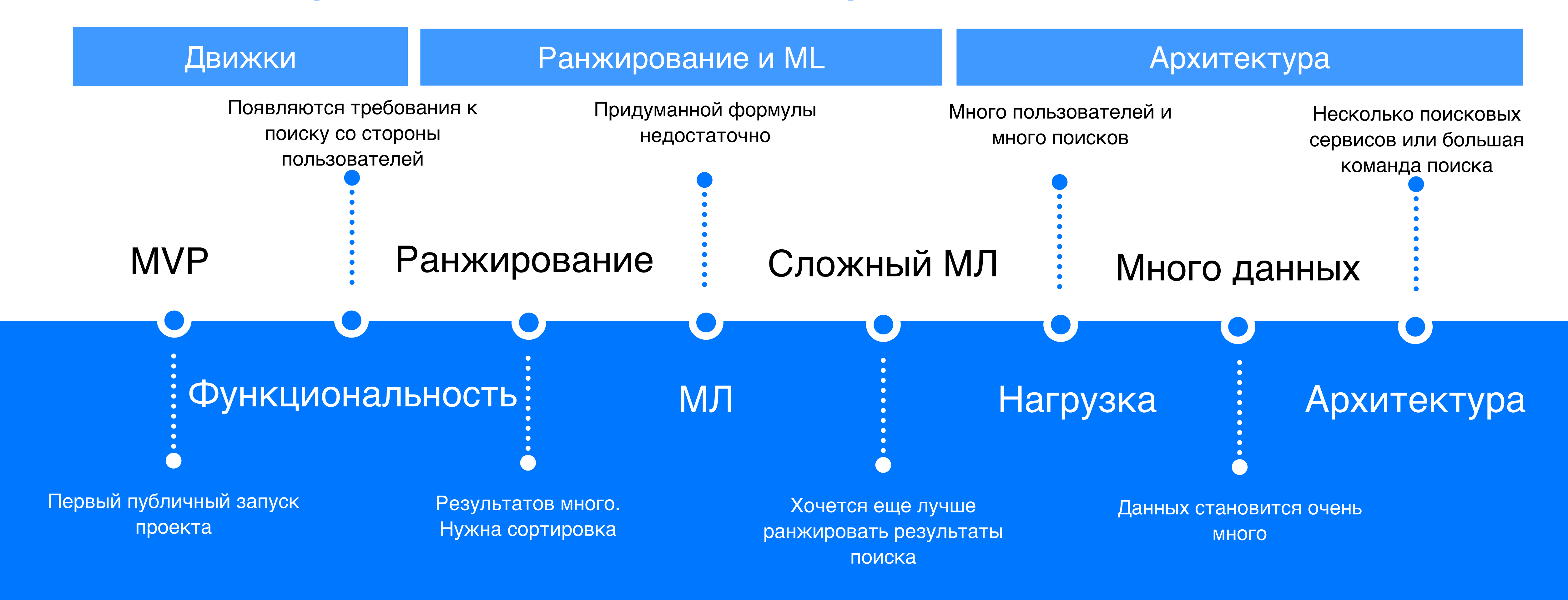
50 млн
поисков в сутки

Возможна ли другая архитектура?

Альтернативная архитектура поиска



Стадия развития вашего проекта





Будем ВКонтакте!

Богдан Гаркушин